Leakage Prevention Method for Unstructured Data Based on Classification

Hao Li[1], Zewu Peng[1], Xinyao Feng[1], and Hongxia Ma[2(✉)]

[1] Information Center, Guangdong Power Grid Corporation, Guangzhou, China
lihao2046@163.com,{pengzewu,fengxinyao}@gdxx.csg.cn
[2] SKLOIS, Institute of Information Engineering, CAS, Beijing, China
mahongxia@iie.ac.cn

Abstract. There is often a lot of sensitive information in the unstructured data of enterprise information network, if not controlled, the sensitive data will flow from the Intranet to the Extranet, which can easily lead to disclosure of corporate information assets, causing serious damage to enterprises. This paper combines the methods of keyword filtering and data label to protect the unstructured data in the corporate data assets based on grading and classification. It is an effective solution to the problem of data leakage, and can greatly reduce false positives in the information protection process, finally improve the accuracy of unstructured data protection. It uses the ElGamal signature algorithm to generate a digital label. Only those people with sensitivity level of keys can make tags with a sensitivity level of document, others cannot replace digital label. At the same time, the network protection server only needs to use the corresponding public key to verify the signature without knowing the private key, thereby effectively ensuring the security of the system.

Keywords: Keyword filtering · Data labels · Signature algorithm

1 Introduction

The enterprise network architecture is generally divided into three parts, that is, the internal network (Intranet), the production Extranet (Extranet) and Internet (Internet). The unstructured data (such as various documents, pictures, etc.) in the enterprise information network is sent from the company's internal network to the Internet or the production Extranet, mainly through a variety of terminal transmission software (such as QQ, Baidu Cloud, mail systems, etc.). Yet there is often a lot of sensitive information in these unstructured data, if such information is not controlled, the sensitive data will flow from the Intranet outside to the Extranet, which can easily lead to disclosure of corporate information assets, resulting in heavy losses to the enterprise [1].

This paper presents the unstructured data assets leakage prevention method based on grading and classification, which combines the methods of keyword filtering and data label to protect the unstructured data in the corporate data assets based on grading and classification. Thereby it is an effective solution to the problem of data leakage, and can greatly reduce false positives in the information protection process, finally improve

W. Niu et al. (Eds.): ATIS 2015, CCIS 557, pp. 337–343, 2015.
DOI: 10.1007/978-3-662-48683-2_30

the accuracy of unstructured data protection. In addition, it uses the ElGamal signature algorithm to generate a digital label. Only those people with sensitivity level of keys (that is, secret-related people) can make tags with a sensitivity level of document, others cannot replace digital label. At the same time, the network protection server only needs to use the corresponding public key to verify the signature without knowing the private key, thereby effectively ensuring the security of the system.

2 Related Work

Sensitive information is contained in the enterprise unstructured data, including documents, images, audio and video materials, so in addition to leakage prevention for structured data, leakage prevention for unstructured data is also an important means of corporate data assets disclosure [2]. Gartner notes that among the five major steps of protecting corporate data assets from loss and information from leaking, the first step is to monitor and filter the content of export network traffic [3, 4]. And its premise is identifying the unstructured data containing sensitive information [5, 6]. However, it is not accurate to use labels with sensitivity level to identify unstructured data containing sensitive information, or to use content filtering method to prevent unstructured data with sensitivity level from leaking, both of which tend to have some false positives. It is clear that today with the network and the big data technology growing, it is not enough to solely rely on the traditional content filtering method to prevent disclosure of unstructured data assets [7]. S.W. Ahmad *et al*. has proposed that currently it is a new research direction of leakage prevention for sensitive data with the cryptographic algorithm method [8]. For unstructured data leakage prevention, Michael Hart *et al*. proposed a text classification method for data loss prevention [9]. Recently, X. Chen *et al*. proposed a cloud security assessment system based on classifying and grading [10].

3 Leakage Prevention Method for Unstructured Data Based on Classification

3.1 Automatic Verification Methods Based on Data Label

In general, he who generates sensitive information is the personnel who matches with the secret level or with a higher level. So from the time when sensitive information is generated, it is added a digital label with secret levels by the secret-related personnel who has produced it, that is, its producer uses a pre-assigned private key adapted to the secret level to sign the document. When the document arrives at the Intranet and Extranet exits, in addition to using keywords filtering method to detect the sensitivity level of the document, the network protection server also needs to use the public key with the secret level to verify the signature. If both are validated, it will indicate that the document is sensitive information, whose request should be blocked immediately.

Specifically, the system first should be three public-private key pairs (sk1, pk1), (sk2, pk2), (sk3, pk3) of high sensitivity level, of sensitivity level and with internal data pre-assigned respectively. For example, when a high sensitive document is produced, the

producer needs to use highly-sensitive private sk1 to have an ElGamal signature on the document and add the signature to the end of the document. When the document reaches the exit of the Intranet, the network protection server will firstly use the keyword filtering method to determine the sensitivity level of the document, and then use the appropriately highly-sensitive public pk1 to verify the signature. If it is verified, the network request will be blocked.

Among them, the digital signature for unstructured data using ElGamal signature algorithm and the verification of it include the steps as follows:

1. Initialization
   Select a large prime number p and a generator Z_p in the controlled terminal, and publish p and g;
   Then select a random number $sk \in Z_{p-1}$, and calculate $pk = g^{sk}(\mathrm{mod}p)$, and disclosure pk as a public key and make sk as a secret key;
2. Sign the document 'm'
   Choose a random number $k \in Z^*_{p-1}$, to calculate $r = g^k(\mathrm{mod}p)$;
   Solve the equation $m \equiv skr + ks(\mathrm{mod}p - 1)$, then get s. In the equation, m is the document needed encrypting; and (r, s) generated after encryption is the signature of the document m, which is attached to the end of the document m;
3. Verification
   Verify the equation $g^m \equiv pk^r r^s(\mathrm{mod}p)$ is right or not, if right then it can be verified.

3.2 Protection Policy of Network Protection Server

Policy = <data type> <sensitivity level of data> <match type> <filter range> <request type> <source IP> <purpose IP> <if the signature is verified> <response action> <severity level>

<Data type> = {. doc, .docx, .txt, .xls, .xlsx, .rar, .wps, .ppt, .pptx, .vsd}
< Sensitivity level of data > = {highly sensitive, sensitive, internal, public}
<Match type> = {case sensitive, case insensitive}
<Filter range> = {cover, subject, body, attachments}
<Request type> = {HTTP, HTTPS, FTP, SMTP}
<Source IP> indicates a device IP sending this information
<Purpose IP> indicates the device IP receiving this information
<Whether the signature is verified> = {Yes, No}
<Response action> = {block, record}
<Severity level> = {high, medium, low, none}

Leakage Prevention Strategy of Highly Sensitive Data. Strategy 1 = <data type = all> <sensitivity level of data = high sensitivity> <match type = case-insensitive> <filter range = all> <request type = any> <source IP = Intranet IP> <destination IP = Extranet IP> <if the signature is verified = yes> <response action = block> <severity rating = high>

Strategy 1 is that it filters the keywords of all data types with case-insensitive, and checks their cover, subject, body and attachments, if it finds the keywords of highly-sensitive, the source IP being the Intranet IP, the destination IP being the Extranet IP, and the use of highly-sensitive public keys to sign the document and get the validation, then it will immediately block any form of the request in HTTP/HTTPS/FTP/SMTP. The event severity rating is high.

Leakage Prevention Strategy of Sensitive Data. Strategy 2 = <data type = all> <data sensitivity level = sensitive> <match type = case-insensitive> <filter range = all> <request type = any> <source IP = Intranet IP> <purpose IP = Extranet IP> <if the signature is verified = yes> <response action = block> <severity rating = high>

Strategy 2 is that it filters the keywords of all data types with case-insensitive, and checks their cover, subject, body and attachments, if it finds the keywords of sensitive, the source IP being the Intranet IP, the destination IP being the Extranet IP, and the use of sensitive public keys to sign the document and get the validation, then it will immediately block any form of the request in HTTP/HTTPS/FTP/SMTP. The event severity rating is medium.

Leakage Prevention Strategy of Internal Data. Strategy 3 = <data type = all> <data sensitivity level = internal> <match type = case-insensitive> <filter range = all> <request type = any> <source IP = Intranet IP> <purpose IP = Extranet IP> <if the signature is verified = yes> <response action = block> < rating = low>

Strategy 3 is that it filters the keywords of all data types with case-insensitive, and checks their cover, subject, body and attachments, if it finds the keywords of internal, the source IP being the Intranet IP, the destination IP being the Extranet IP, and the use of internal public keys to sign the document and get the validation, then it will immediately block any form of the request in HTTP /HTTPS/FTP/SMTP. The event severity rating is low.

Leakage Prevention Strategy of Public Data. Strategy 4 = <data type = all> <data sensitivity level = public> <match type = case-insensitive> <filter range = all> <request type = any> <source IP = Intranet IP> <purpose IP = Extranet IP> <response action = register> <severity rating = none>

Strategy 4 is that it filters the keywords of all data types with case-insensitive, and checks their cover, subject, body and attachments, if it finds the keywords of public, the source IP being the Intranet IP, and the destination IP being the Extranet IP, then it will record any form of the request in HTTP/HTTPS/FTP/SMTP.

3.3 The Implementation of Unstructured Data Assets Leakage Prevention Method Based on Classification

It comprises the following steps as in Fig. 1:

1. The controlled terminal makes classification and grading of unstructured data assets, and makes digital signatures of the unstructured data based on sensitivity levels according to corresponding types, including the following steps: The controlled

terminal makes classification and grading of unstructured data assets, and divides them into data of highly-sensitive level, sensitive level, internal level and public level. Then it pre-assigns public and private key pairs to data of highly-sensitive level, sensitive level and internal level respectively, and uses each data to the corresponding private key to make ElGamal signature. The specific steps of using ElGamal signature algorithm to make digital signatures for highly sensitive, sensitive and internal levels of unstructured data, and verifying the signatures, please refer to Sect. 3.1.

2. When the controlled terminal requests to send unstructured data to the Internet or Extranet, the network protection server will filter the data with mirror traffic and sensitive keywords. It includes such details as follows: When the controlled terminal issues the requests of HTTP, HTTPS, FTP or SMTP to send unstructured data to the Internet or the Extranet, the network protection server will judge - if the source IP is the Intranet IP, and the destination IP is the Extranet IP, then it will filter the cover, subject, body and attachments of the unstructured data with mirror traffic and sensitive keywords, to determine if it contains sensitive keywords.
3. If the unstructured data contains sensitive keywords, then use the public key of the corresponding sensitivity level to verify the signature of the unstructured data;
4. If it is verified, then it will block the request of the controlled terminal to send data to the Internet or to the Extranet.

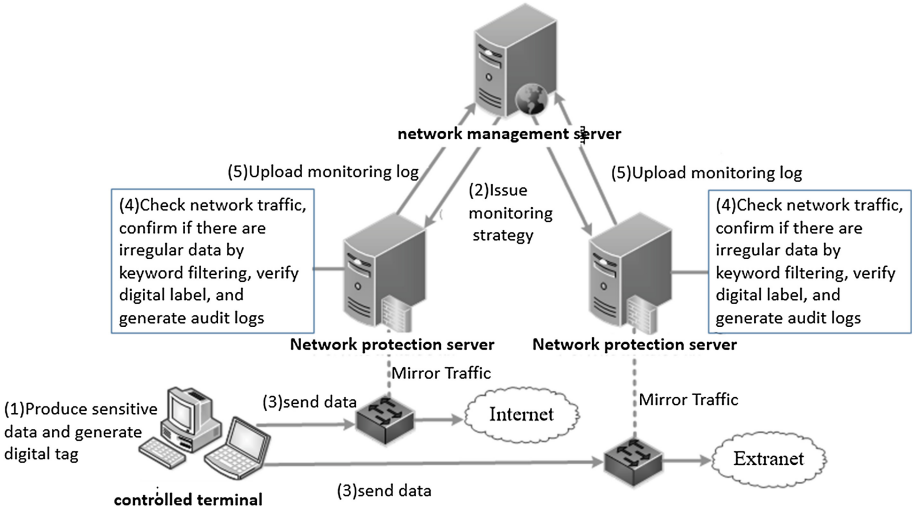

Fig. 1. Hardware connection structure and work principle sketch map in the information leakage protection system

4 Conclusion

There is often a lot of sensitive information in the unstructured data of enterprise information network, if not controlled, the sensitive data will flow from the Intranet to the Extranet, which can easily lead to disclosure of corporate information assets, causing serious damage to enterprises. This paper combines the methods of keyword filtering and data label to protect the unstructured data in the corporate data assets based on grading and classification. It effectively solves the problem of data leakage, and greatly reduces false positives in the information protection process (such as taking the non-sensitive information sensitive information), finally improve the accuracy of unstructured data protection.

In addition, if you use the Hash algorithm in the existing technology to generate a digital label, as long as you know the Hash algorithm, anyone can generate and verify the Hash value, and for the same document, the Hash value generated is the same, then it cannot guarantee that the documents with sensitive level- can only be produced by a person with the corresponding secret level, and anyone can modify the document to re-generate Hash value, therefore it is not conducive to the system safety. Only those people with sensitivity level of keys (that is, secret-related people) can make tags with a sensitivity level of document, others cannot replace digital label. At the same time, the network protection server only needs to use the corresponding public key to verify the signature without knowing the private key, thereby effectively ensuring the security of the system.

Acknowledgement. This work was sponsored by the Information Center of Guangdong Power Grid Corporation's project of Study on Data Security in Big Data Environments (No. K-GD2014-1019) and Xinjiang Uygur Autonomous Region science and technology plan (No. 201230121), the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDA06040601).

References

1. Shabtai, A., Elovici, Y.: A Survey of Data Leakage Detection and Prevention Solutions. SpringerBriefs in Computer Science. Springer, Heidelberg (2012)
2. Tankard, C.: Big data security. Netw. Secur. **2012**(7), 5–8 (2012)
3. Mogull, R.: Top five steps to prevent data loss and information leaks. Gartner Research, 12 July 2006
4. Ouellet, E., McMillan, R.: Magic quadrant for content-aware data loss prevention. Gartner Research, 10 August 2011
5. Byers, A.C., Renfro, C., Pendleton, C., et al.: Method for analyzing and managing unstructured data. US Pattern. US8122510 B2. 2012.2.21
6. Geethakumari, G., Srivatsava, A.: Big data analysis for implementation of enterprise data security. Int. J. Comput. Sci. Inf. Technol. Secur. **2**(4), 742–746 (2012)
7. Besser, S.: Stopping information leaks: Why traditional content filtering is no longer enough. White paper of Port Authority Technologies (2005)
8. Ahmad, S.W., Bamnote, G.R.: Data leakage detection and data prevention using algorithm. Int. J. Comput. Appl. **6**(2), 394–399 (2013)

9. Chen, X., Chen, C., Tao, Y., Hu, J.: Cloud security assessment system based on classifying and grading. IEEE Cloud Comput. Mag. **2**(2), 58–67 (2015)
10. Hart, M., Manadhata, P., Johnson, R.: Text classification for data loss prevention. In: Fischer-Hübner, S., Hopper, N. (eds.) PETS 2011. LNCS, vol. 6794, pp. 18–37. Springer, Heidelberg (2011)